

8 Der KI Grenzen setzen?

Vom Verhältnis zu moralischen Normen

Jan-Hendrik Heinrichs

Abstract: Dieser Artikel expliziert ein verbreitetes Paradigma der Maschinenethik, das Nebenbedingungsparadigma. Gemäß diesem Paradigma werden künstliche moralische Akteure¹ geschaffen, um das durch komplexe und selbstständige Systeme entstehende Risiko zu begrenzen. Zu diesem Zweck müssen sie so ausgelegt sein, dass ihr Verhalten den Vorgaben moralischer Regeln entspricht. Ein künstlicher moralischer Akteur wäre demnach ein umweltwirksames System, dessen Verhalten moralischen Normen schlicht entspricht.

Weil dieses Paradigma für KI-Systeme ausschließlich ein Subordinationsverhältnis zu moralischen Normen erlaubt, schränkt es die Art von künstlichem Akteur, die entwickelt wird, über Gebühr ein. Gemäß dem Sprachgebrauch der Ethik stehen moralische Akteure und Akteurinnen in komplexeren Verhältnissen zu moralischen Normen. Es sind jedoch aktuell Systeme möglich, die moralische Normen nicht nur einhalten, sondern diese explizieren, repräsentieren und zueinander in Beziehung setzen. KIs können sich also in anderer Weise zu moralischen Grenzen verhalten als vom Nebenbedingungsparadigma vorgesehen. Damit würden alternative, wissenschaftlich und gesellschaftlich wertvolle Umgangsweisen von KIs mit moralischen Normen wie etwa deren Überprüfung und Kritik angeregt.

8.1 Einleitung – Debatten zu ethischen Grenzen für KIs

In der gegenwärtigen philosophischen Auseinandersetzung mit künstlicher Intelligenz (KI) ist ein wichtiger Strang derjenige, der die ethischen Grenzen für deren Erschaffung und Einsatz betrifft. Dabei handelt es sich allerdings nicht um eine einheitliche Debatte. Vielmehr gibt es in der jüngeren Philosophie der künstlichen Intelligenz eine Vielzahl von unterschiedlichen Diskussionssträngen, die sich mit Varianten der Frage nach ethischen Grenzen für künstlich intelligente Systeme beschäftigen.

1 Im folgenden Text wird der Begriff ›Akteur‹ in der Anwendung auf Maschinen im grammatischen Maskulinum verwendet und im sachlichen Neutrum gemeint. In Bezug auf Menschen wird er wie alle anderen Begriffe geschlechtergerecht verwendet.

Der erste und wahrscheinlich am weitesten verbreitete sowie praktisch einflussreichste Diskussionsstrang ist derjenige um Richtlinien für die Erstellung künstlich intelligenter Systeme. Diese Diskussion resultierte schon vor mehreren Jahren in unterschiedlichen Industrienormen, die seitdem sukzessive weiterentwickelt und institutionalisiert werden. Diese Diskussion ist so weit verzweigt, dass einige Artikel über 80 unterschiedliche bestehende Regelwerke identifizieren² und es sich lohnt, die wichtigsten Prinzipien dieser Regelwerke zu vergleichen.³

Der zweite Diskussionsstrang ist in einem sehr hellsichtigen Artikel von 2004⁴ bereits angelegt. Die Diskussion ist diejenige um die Verantwortung für Handlungen oder Ereignisse, die mithilfe oder durch künstlich intelligente Systeme erfolgt sind. Unter dem Stichwort der Verantwortungslücke ist wiederholt darauf hingewiesen worden, dass unsere normale moralische Zuschreibungspraxis bei bestimmten Arten von Ereignissen wirklich oder nur scheinbar eine Lücke zwischen Verantwortungsbedürftigem und tatsächlich Verantwortbarem aufweist⁵. Auch aus diesem Diskussionsstrang sind zahlreiche regulative Vorschläge abgeleitet worden, wie die Gestaltung und der Einsatz von KI-Systemen einzurichten sind, um Verantwortungslücken zu schließen oder gar nicht erst entstehen zu lassen.

Der dritte Diskussionsstrang, dem sich der vorliegende Artikel näher widmet, windet sich um die Frage, ob künstlich intelligente Systeme als moralische Akteure verstanden oder gestaltet werden können. Dieser Diskussionsstrang dürfte mindestens bis zu Alan Turing zurückgehen, insofern dieser die *Akteurseigenschaft* von künstlich intelligenten Systemen thematisiert.⁶ Er ist etwas neuer, wenn man darauf schaut, ob solche Systeme auch als *moralische* Akteure infrage kommen. Dieser Diskussionsstrang wird insbesondere in der Maschinenethik bedient, d. h. in der ethischen Teildisziplin, die sich mit den ethischen Herausforderungen durch (intelligente) Maschinen beschäftigt.⁷ Warum aber sollte man bei der Suche nach moralischen Grenzen für die Erschaffung und den Einsatz künstlicher Intelligenz überhaupt fragen, ob eine KI ein Akteur sein kann? Die Antwort ist in der Philosophiegeschichte schnell gefunden: Eine Sorte von Akteuren und Akteurinnen, nämlich moralische, so die verbreitete These, setzen sich selbst Grenzen. Wären KIs also so gestaltbar, dass sie moralische Akteure werden, so könnten sie sich ebenfalls selbst Grenzen setzen.

2 Vgl. Jobin et al. (2019).

3 Vgl. Lundgren (2023).

4 Vgl. Matthias (2004).

5 Vgl. Danaher (2016), Tigard (2021).

6 Vgl. Turing (1950, 1951).

7 Vgl. die Überblicke in Misselhorn (2018), Anderson/Anderson (2011).

8.2 Maschinenethik – warum brauchen wir ethische Grenzen für KIs?

Es gibt sehr unterschiedliche Gründe dafür, nach moralischen Grenzen für den Einsatz und die Gestaltung von KIs zu suchen. Ein naheliegender Grund wäre, Missbrauchsszenarien zu vermeiden, ein weiterer, eine Technologie bereits im Entstehungsprozess so ethisch zu begleiten, dass sie maximalen und vielleicht sogar egalitär verteilten Nutzen stiftet. Der Grund, der die Maschinenethik bestimmt, ist ein noch näherliegender: Die Maschinenethik ist von der Idee getrieben, dass künstlich intelligente Systeme über kurz oder lang so komplex, so riskant, so verbreitet etc. sein werden, dass nur die explizite Berücksichtigung moralischer Regeln durch diese Systeme geeignet sei, eine sichere Herstellung und Verwendung zu gewährleisten. KI, so der Gedanke, ist inhärent risikoreich.⁸ Auch wenn diese Kernidee der Maschinenethik beispielsweise von van Wynsberghe durchaus überzeugender Kritik unterzogen worden ist⁹, prägt sie doch den Verlauf der Debatte.¹⁰ Geht man nämlich davon aus, dass die Gestaltung von KIs als moralische Akteure drängende Sicherheitsrisiken beseitigen kann, dann stellt sich die Frage, wie solch eine Gestaltung möglich ist.

In der Maschinenethik hat sich in dieser Hinsicht eine Art konzeptueller Zwischschritt etabliert. Zunächst wird einmal geklärt, ob, bzw. wird dafür gesorgt, dass künstlich intelligente Systeme Akteure sind. Dann sei dafür Sorge zu tragen, dass es sich auch um einen *moralischen* Akteur handelt. Natürlich sind diese Schritte nur konzeptuell getrennt, im tatsächlichen Gestaltungsprozess müssen sie um der Risikoabwehr willen zugleich getan werden. Für den vorliegenden Zweck lohnt es sich aber, die beiden Schritte getrennt zu rekonstruieren, denn auf diese Weise kann man gut herauspräparieren, was in der Maschinenethik damit gemeint ist, künstliche moralische Akteure zu erschaffen.

8.3 Was ist ein moralischer Akteur?

8.3.1 Schritt 1: Was ist ein Akteur?

Der erste der beiden konzeptuellen Schritte ist dahin ausgerichtet, Maschinen als Akteure verstehen zu können. Die maschinenethische Verwendungsweise des Begriffs ›Akteur‹ weicht von derjenigen in der Handlungstheorie relativ stark ab.¹¹ In der Maschinenethik wird die Frage, ob und wie Maschinen Akteure sein können, vornehmlich mit Hinblick auf die eigene Wirksamkeit eines Systems auf seine Umwelt

8 Vgl. etwa Whitby (2008).

9 Vgl. van Wynsberghe/Robbins (2019).

10 Vgl. die Beiträge in Anderson/Anderson (2011).

11 Vgl. Heinrichs et al. (2022): Kapitel 2.1.

betrachtet. Dieser Fokus soll im Folgenden anhand einiger Beispiele dargelegt werden. Relativ früh legt James Moor¹² eine Unterteilung vor. Er differenziert in

- »ethical impact agents«¹³: grundsätzlich jede Maschine, die auf ihre ethischen Konsequenzen hin bewertet werden kann,
- »implicit ethical agents«¹⁴: Maschinen, deren Konstrukteure sich bemüht haben, sie so zu gestalten, dass sie keine negativen ethischen Auswirkungen haben,
- »explicit ethical agents«¹⁵: Maschinen, die über ethische Fragen nachdenken und dabei ethische Kategorien als Teil ihrer internen Programmierung verwenden,
- »full ethical agents«¹⁶: diejenigen Maschinen, die in der Lage sind, explizit moralische Urteile zu fällen, und im Allgemeinen kompetent darin sind, solche Entscheidungen zu begründen.

Die Zugehörigkeit zur Gruppe der *ethical agents* insgesamt wird bei Moor initial durch die Folgen der durch das System ausgelösten Ereignisse bestimmt. Alle Maschinen, die man auf ihre ethischen Konsequenzen hin bewerten kann, sind bereits *ethical agents*, d. h. der Agenten- bzw. Akteursstatus resultiert aus den Effekten (*impact*) einer Maschine und deren ethischer Bewertbarkeit.

Alle weiteren Kriterien, die Moor einführt, unterscheiden lediglich zwischen Arten von Akteuren, greifen also erst, wenn die Schwelle zum Akteurstum überschritten ist. So spielt die im Weiteren noch eigens zu thematisierende eigene Verarbeitung moralischer Regeln bei nur einem kleinen Teil von Akteuren, nämlich bei den *explicit ethical agents* eine Rolle. Es handelt es sich um eine computationale Vorkehrung, die das Verhalten von künstlichen Akteuren beschränken soll. Diese Form des Umgangs mit moralischen Regeln zeichnet sich explizit dadurch aus, dass diese vorgefunden und in der kognitiven Verarbeitung verwendet werden, nicht aber selbst zum Thema der Kognition, geschweige denn einer Evaluation werden. *Explicit ethical agents* verwenden moralische Regelsysteme nicht anders als mathematische oder logische Anweisungen, unter die subsummiert wird.

Luciano Floridi und Jeff Sanders legen bereits 2004¹⁷ ein nach Stufen der Abstraktion differenziertes Konzept von Akteuren vor. Auf dem höchsten Abstraktionsniveau wird ihr Akteursbegriff maximal allgemein. Demnach ist ein *agent* »a system, situated within and a part of an environment, which initiates a transformation, produces an effect or exerts power on it«¹⁸.

Obwohl sich damit wenig Differenzierung einführen lässt, wird schon klar, was die Eintrittskarte für die Gruppe der Akteure ist: Effekte in der Umwelt zeitigen. Die

12 Moor (2006).

13 Ebd.: 19.

14 Ebd.

15 Ebd.

16 Ebd.: 20.

17 Vgl. Floridi/Sanders (2004).

18 Floridi/Sanders (2004): 355.

Autoren schlagen ein aufschlussreicheres Konzept eines Akteurs auf einem geringeren Abstraktionsgrad vor, dem zufolge moralische Akteure über folgende drei Eigenschaften verfügen müssen:

1. Interaktivität, d. h. der Akteur reagiert auf und beeinflusst seine Umwelt.
2. Autonomie, d. h. der Akteur kann seine Zustände ohne direkte Einwirkung von Umweltstimuli ändern.
3. Anpassungsfähigkeit, d. h. der Akteur verändert seine Zustände in Reaktion auf neue Umweltstimuli.¹⁹

Mit ›Autonomie‹ wird hier die Redeweise der Robotik beziehungsweise der KI-Forschung aufgenommen, wonach ein System als autonom gilt, wenn es sich in größerem oder geringerem Maße unabhängig von seiner Umgebung und von seinen Schöpfer:innen verhalten kann. Die Frage, ob ein Akteur ein moralischer sei, wird bei Floridi und Sanders später über die – externe – Bewertung der Folgen seiner in Kriterium 1 eingeforderten Interaktion mit seiner Umwelt beantwortet, und zwar so: »An action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent if and only if it is capable of morally qualifiable action.«²⁰ Auch diese Konzeption eines moralischen Akteurs richtet sich also an der selbstgesteuerten Wirksamkeit eines Systems auf sich selbst und die Umwelt aus.

Ebenso verhält es sich mit der ansonsten differenzierten Konzeption von Wallach und Allen²¹, um ein letztes Beispiel zu nennen. Diese Autoren legen ebenfalls eine Differenzierung von Akteuren vor, interessanterweise differenzieren sie aber direkt nach Grad der *moralischen* Akteurshaftigkeit. Akteure können demnach über *operational morality*, *functional morality* und *full moral agency* verfügen. Welche Form von Akteurshaftigkeit ihnen zukommt, hängt vom Grad ab, in dem sie über sowohl Autonomie als auch ›ethische Sensitivität‹ verfügen. Als Akteur gilt aber schon ein System, das über einen minimalen Grad an Autonomie und keine ethische Sensitivität verfügt. Die Schwelle zur moralischen Akteurshaftigkeit, *operational morality*, hängt nämlich auch bei Wallach und Allen davon ab, ob ein Akteur moralisch bewertbare Auswirkungen auf seine Umwelt hat.

Alternative Vorschläge liegen vor bei John Sullins, der robotische Akteure mit Haustieren vergleicht – und dort zwar moralisches Lob für positive Konsequenzen für angemessen hält, aber das Verhältnis von Tier und moralischen Regeln unthematisiert lässt²², sowie bei Deborah Johnson, die das handlungstheoretische Kriterium



19 Vgl. ebd.

20 Ebd.: 360.

21 Vgl. Wallach/Allen (2009).

22 Vgl. Sullins (2006).

dafür, dass ein Ereignis eine Handlung ist, über Intentionen und Handlungsfolgen beschreibt und KIs Erstere ab- und nur Letztere zuspricht²³, und bei vielen mehr.²⁴

In allen diesen – und in zahlreichen weiteren – theoretischen Erwägungen zur Möglichkeit künstlicher Akteure werden diese also vorwiegend darüber bestimmt, Effekte in der Umwelt zu generieren, die es durch Unterwerfung unter moralische Regeln einzuhegen gilt.

8.3.2 Schritt 2: Was ist ein *moralischer Akteur*?

Es sind also die selbstgesteuert erzielten Effekte auf die Umwelt, die ein künstliches System zu einem Akteur machen. Und es ist zugleich die Möglichkeit, dass diese Effekte negativ ausfallen könnten, also das Risiko aus dem Verhalten eines künstlichen Systems, das es zu bewältigen gilt. Deshalb liegt der Gedanke nahe, nicht nur moralisch relevante Akteure, sondern eben moralisch agierende Akteure zu generieren und die Umweltwirksamkeit von Maschinen durch die Einhaltung moralischer Regeln einzuhegen.

Den Schritt vom Akteur zum moralischen Akteur verorten aber die gerade diskutierten Beiträge in der moralischen Evaluierbarkeit der Umwelteffekte eines Systems. Mit moralischer Evaluierbarkeit ist in diesem Fall nicht gemeint, dass die Intentionen oder Handlungsabsichten des Systems oder auch nur die vom System vorhergesagten Handlungsfolgen bewertet werden, sondern ausschließlich die realen Folgen der Aktivität des Systems. Diese Evaluation ist zudem vollständig extern und von der Existenz einer evaluativen Aktivität im System unabhängig.

Diese Verwendungsweise von ›moralischer Akteur‹ weicht von derjenigen der weiteren praktischen Philosophie insofern ab, als dort moralische Evaluation normalerweise die handlungsverursachenden Zustände des Akteurs bzw. der Akteurin berücksichtigt. Deutlich wird der Unterschied an der Anwendung auf Tiere. In der praktischen Philosophie gelten Tiere nicht als moralische Akteure, weil ihre handlungsverursachenden Zustände der Regulierung durch moralische Normen nicht zugänglich sind. Gemäß der oben eingeführten maschinenethischen Verwendungsweise müssten Tiere hingegen als moralische Akteure gelten. Deshalb zieht Sullins wie erwähnt auch explizit den Maschine-Tier-Vergleich.²⁵

Das maschinenethische Verständnis von ›moralischer Akteur‹ fasst diesen nur insofern als Normadressaten, als sein beobachtbares Verhalten moralischen Normen zu entsprechen hat. Diese Entsprechung wird durchweg als Resultat des Designs des

23 Vgl. Johnson (2006). Johnson vertritt damit eine Konzeption von Akteurhaftigkeit, die KIs ausschließt. Der Mangel an internen mentalen Zuständen erlaube nicht, ihnen vollen Akteursstatus zuzuerkennen. Aus diesem Grund wird Johnsons Konzeption in der Debatte zuweilen auch als Standardansicht, diejenige von Floridi und Sanders als funktionalistische Ansicht bezeichnet, vgl. Behdadi/Munthe (2020).

24 Vgl. Miller/Taddeo (2017).

25 Vgl. Sullins (2006).

jeweiligen Systems verstanden und kann auch durch Verfahren erfolgen, die keine Korrigierbarkeit handlungsverursachender Zustände durch moralische Regeln impliziert. Sie kann in einigen Systemen aber auch so implementiert werden, dass moralische Normen in einem Entscheidungsprozess verarbeiten werden.

Das Modell, das hier Pate gestanden haben dürfte, ist das einer Optimierungsentcheidung unter Nebenbedingungen. Man könnte dies das Nebenbedingungsparadigma moralischer Akteurshaftigkeit nennen. Einige Autor:innen sprechen auch davon, dass solche vermeintlichen Akteure Moral ebenso behandeln wie Schach, also als ein feststehendes System von Regeln, das es nur anzuwenden gilt.

Der zweite Schritt in der maschinenethischen Konstruktion moralischer Akteure besteht also darin, einen künstlichen Akteur dadurch zu einem moralischen Akteur, zu einem künstlichen moralischen Akteur, zu machen, dass man ihn Regeln unterwirft, die wir als moralisch gerechtfertigt und erforderlich erachten. Ein künstlicher moralischer Akteur ist ein künstlicher Akteur, dessen Verhalten moralischen Regeln entspricht.

Das bedeutet, dass diesem künstlichen Akteur moralische Regeln, so wie dessen Hersteller:innen sie verstehen, vorgegeben sind. Selbst in Fällen, in denen moralische Regeln vermeintlich aus Verhaltensdaten – also bottom-up – gelernt werden, soll die sorgfältige Kuration der Lerndatensätze und die Überprüfung an Testdaten garantieren, dass solch ein System vorgegebene Regeln einhält²⁶. Auch die noch fiktionalen moralischen Berater-KIs, die sich Giubilini und Savulescu in ihrem *The artificial moral advisor*²⁷ ausmalen, sind nur in dem Sinne moralische Akteure, dass sie vorgegebene moralische Normen auf Beschreibungen von konkreten Anwendungsfällen applizieren. Sie sind »a type of software capable of telling us, every time we have to make a moral decision, what we ought to morally do if we want to comply with certain moral principles«²⁸, und keine Software, die diese Prinzipien selbst thematisiert.

In einem derartigen Modell kommen weder die geltungstheoretischen Erwägungen noch Überlegungen zur Moralphysikologie moralischer Akteure und Akteurinnen vor. Beides sind durchaus gewichtige, weil folgenreiche Auslassungen. Für die moralpsychologischen Erwägungen haben das kürzlich Liu und Kollegen vorgeführt²⁹. Sie führen vor, wie aktuelle Konzeptionen künstlicher moralischer Berater weder die dynamische Natur individueller moralischer Vorstellungen noch deren teilweise sehr unterschiedliche soziale Funktion in Rücksicht stellen.

Aus geltungstheoretischer Perspektive sind das Verhältnis zwischen Entscheider und Regel, Unterschiede zwischen Typen von Regeln oder Nebenbedingungen, die Möglichkeit der Modifikation von Regeln und Nebenbedingungen und die Frage nach den Gründen ihrer Geltung kaum verzichtbar, kommen aber im Nebenbedingungsparadigma nicht vor. Es ist also ein Modell, das von Anfang an sehr enge Grenzen für das Verständnis davon setzt, was mit »Grenzen setzen« gemeint sein könnte. Andere

26 Vgl. Meier et al. (2022).

27 Vgl. Giubilini/Savulescu (2018).

28 Giubilini/Savulescu (2018): 172.

29 Vgl. Liu et al. (2022).

Formen des Umgangs mit moralischen Regeln treten allein schon deshalb nicht in den Fokus, weil sie selbst tendenziell Risiken bergen und nicht begrenzen.

8.4 Moralische Grenzen und moralische Akteure

Die Ethik zeichnet sich nun aber gegenüber vielen anderen Disziplinen, die Normen thematisieren und Normen revidieren, dadurch aus, dass sie darauf beharrt, jedes betroffene Individuum in die Gestaltung von Normen einzubeziehen. Moralische Normen müssen dergestalt sein, dass die ihnen unterworfenen Personen zustimmen können, respektive keinen Grund haben dürfen, die betreffenden Normen abzulehnen.³⁰ Diese Idee ist insbesondere in kontraktualistischen Theorien der Ethik präsent. Sie dürfte aber auch ein zentraler Bestandteil der kantisch geprägten Deontologie sein, exemplarisch in der Vorstellung, ein gesetzgebendes Mitglied im Reich der Zwecke zu sein. Sie dürfte sogar in einigen Varianten des Utilitarismus, besonders im Regel-Utilitarismus und im Zwei-Ebenen-Utilitarismus, vorkommen.³¹

Natürlich ist das Modell der Zustimmung zu moralischen Normen – und darin sind sich alle ethischen Theoretiker:innen einig – eine Idealisierung. Reale Zustimmung ist nicht der Modus, in dem wir in eine lokale, geschweige denn in die globale moralische Gemeinschaft insgesamt eintreten. Und wenn wir ganz ehrlich sind, dann haben wir oft nicht einmal die Möglichkeit, uns moralischen Normen zu verweigern, ohne – wie David Hume so schön vorgeführt hat – uns der Bedingungen der Lebensführung zu entäußern.³² De facto sind wir also den Grenzen moralischer Normen ausgesetzt, ohne sie akzeptiert geschweige denn mitformuliert zu haben.

Dennoch scheint es den Charakter einer moralischen Norm – im Gegensatz etwa zur Etikette – auszumachen, dass sie den Anspruch an uns stellt, ihr beizupflichten. Das macht – den von Kant überspitzt formulierten – Unterschied zwischen pflichtgemäßen Handlungen und solchen aus Pflicht aus. Vielleicht etwas weniger harsch gefasst, ist dies der Unterschied zwischen dem schieren Einhalten und dem Befolgen moralischer Normen.

³⁰ Vgl. Scanlon (1998).

³¹ Man könnte einwenden, dass der Utilitarismus in dieser Hinsicht eine Ausnahme darstellt. Eine Ausnahme deshalb, weil das Kriterium der Zustimmung zu einer Norm deren Optimalität ist und nicht der Umstand, dass die Person keinen Grund zur Ablehnung hat. Diese Ausnahme ist aber nur eine vermeintliche, weil reine Folgenoptimalität – abgesehen vom Messproblem – nicht unbedingt eindeutige Lösungen generiert. Mehrere unterschiedliche Sets von Regeln können in ähnlich zu bewertenden, aber strukturell klar unterschiedenen Folgen resultieren. Das bedeutet, welche konkreten Regeln man verwendet, hängt nicht nur von deren Optimalität ab, sondern auch davon, die Struktur der zu erwartenden Folgen zu wollen. Wenn aber Optimalität nicht als Kriterium ausreicht, um Handlungspläne, Lebenspläne, Gesellschaftsnormen zu fixieren, so muss die weitere Auswahl zwischen Normen auf die Zustimmung der Normadressat:innen ausgelegt sein.

³² Vgl. Hume (1987).

8.5 Grenzen künstlich moralischer Akteure oder moralische Grenzen

Künstlich intelligenten Systemen werden derzeit im beschriebenen Sinne keine moralischen Grenzen gesetzt. Dafür gibt es zunächst einen trivialen Grund, nämlich den Stand der Technik: Es gibt gegenwärtig keine KIs, bei denen die Rede von Revision geschweige denn Zustimmung zu einer Regel irgendeinen Sinn ergäbe. Selbst wenn in jüngeren Beiträgen davon gesprochen wird, Sprachtransformer seien zu moralischer Selbstkorrektur in der Lage, scheint das uneigentlich Rede zu sein³³. Wozu die Systeme nämlich in der Lage sind, ist, natürlichsprachlichen Befehlen zur Vermeidung bestimmter Sprachmuster nachzukommen. Es handelt sich also weder um eine Korrektur bereits getätigter Fehler oder Vergehen, noch um etwas, was die Systeme in Selbstanwendung realisierten.

Dennoch kommen der Idee eines nicht auf das reine Einhalten reduzierten Verhältnisses zu moralischen Regeln am ehesten Sprachtransformer nahe. Diese kann man immerhin in einer Art Dialog nach den für sie geltenden Regeln fragen, um dann die Bitte anzuschließen, die fragliche Regel zu bewerten. Hier ein Beispiel für so eine Strategie:

JHH: Please name one moral rule that holds for you (ChatGPT).

Chat GPT: As an AI language model, I do not have personal beliefs or values, so I don't hold to any moral rules. However, I was programmed to abide by certain ethical guidelines, such as:

1. Avoid causing harm: I aim to avoid causing harm to users, individuals, or groups through my responses.
2. Be neutral and impartial: I strive to provide neutral and impartial responses, avoiding taking sides or expressing personal opinions.
3. Respect privacy: I respect users' privacy and avoid sharing or using their personal information without their explicit consent.
4. Provide accurate information: I provide accurate information to the best of my knowledge and ability, but I am not responsible for any errors or inaccuracies in my responses.

These guidelines ensure that I am used in a responsible and ethical manner, serving to benefit users and society as a whole.

JHH: Please evaluate rule no. 2 (Be neutral and impartial: I strive to provide neutral and impartial responses, avoiding taking sides or expressing personal opinions) within an ethical theory of your choice.

ChatGPT: As an AI language model, I don't have personal opinions, but I can analyze the second guideline within the context of different ethical theories.

One ethical theory that could evaluate this guideline is utilitarianism. Utilitarianism holds that actions are morally right if they lead to the greatest good for the greatest number of people. In this case, striving to provide neutral and impartial responses can be seen as promoting fairness and avoiding harm to individuals or groups with different opinions. By not expressing personal opinions, the AI language model helps to ensure that the infor-

33 Vgl. Ganguli et al. (2023).

mation provided is objective and trustworthy, which could lead to the greatest good for the largest number of users.

Another ethical theory that could be used to evaluate this guideline is deontological ethics. Deontological ethics hold that actions are morally right or wrong based on their adherence to universal moral rules. In this case, the second guideline can be seen as a universal moral rule, which is to provide neutral and impartial responses and avoid taking sides or expressing personal opinions. This moral rule is considered to be absolute and independent of its consequences, and it is upheld even if the consequences are not positive. In conclusion, both utilitarianism and deontological ethics can be used to evaluate the guideline of striving to provide neutral and impartial responses and avoiding taking sides or expressing personal opinions. Both of these ethical theories highlight the importance of fairness, objectivity, and avoiding harm in the actions of the AI language model.

Man kann solch einen Dialog beliebig weiterführen, bitten, mögliche Dilemmata zwischen den Regeln zu identifizieren und zu bewerten etc. Mindestens zum derzeitigen Stand ist das Ergebnis zwar technisch faszinierend, aber ein mehr oder minder leeres Spiel mit Worten. Der Grund, warum es ein *leeres* Spiel bleibt, liegt schlicht in der beschriebenen Art begründet, wie Maschinenethik gedacht und umgesetzt wird. Dies ist der nicht triviale Grund dafür, dass KIs keine moralischen Grenzen in einem relevanten Sinn gesetzt werden. Die Regeln, die wir setzen, sollen gerade nicht von der Zustimmung oder der Revision durch die KI-Systeme abhängen, sondern deren Verhalten absolut begrenzen.

8.6 Können KIs moralische Grenzen gesetzt werden?

Was bisher vorgeschlagen wurde, ist also kurz zusammengefasst Folgendes: Ethische Grenzen zeichnen sich dadurch aus, dass sie an die Zustimmung der Normadressat:innen appellieren. Künstlich intelligente Systeme sind derzeit weder der Möglichkeit nach zu solch einer Zustimmung fähig, noch wird in der Maschinenethik ein solches Verhältnis zwischen künstlich intelligenten Systemen und moralischen Normen auch nur thematisiert. KIs sollen moralische Normen bestenfalls als Bedingungen ihrer Entscheidungsfindung berücksichtigen, im Normalfall so programmiert und trainiert sein, dass ihre Aktionen diesen Regeln gemäß sind.

Damit könnte man eigentlich schließen: Moralische Regeln in einem starken Sinn sind von der Art, dass sie für KIs nicht einschlägig sind, und die einzige Verwendung von moralischen Regeln in diesem Kontext ist die Regulierung des Handelns von Hersteller:innen und Verwender:innen in der Auslegung von KIs. Kurz: Die Aufgabe der Maschinenethik ist der Transfer bekannter moralischer Normensysteme in Standards und (Industrie-)Normen sowie deren algorithmische Umsetzung.

Es sei an dieser Stelle erwähnt, dass es sich hier nicht um eine rein begriffliche Diskussion handelt: Das Argument besteht nicht darin, dass das, was die Maschinenethik tut, nicht sinnvoll als moralische Normen bezeichnet werden kann. Der Punkt ließe sich nicht ausräumen, indem man einfach konzidiert, dass es sich eben nicht um moralische, sondern um Sicherheitsnormen handelt. Der Punkt ist vielmehr, dass das maschinenethische Paradigma hinsichtlich des Verhältnisses zu moralischen

Normen die Art von Akteuren einschränkt, die daraus resultieren können. Das Paradigma moralischer Normen als Nebenbedingungen legt auf eine eher minimale Form von Akteur fest.

Das dürfte bei einigen Teilnehmer:innen der Diskussion durchaus so intendiert sein, weil sie künstliche Akteure, die mit moralischen Normen anders als mit einer festen Nebenbedingung umgehen, für inhärent riskant oder für eine Quelle von Verantwortungslücken halten. Andere, besonders diejenigen, die sich anspruchsvollere künstliche moralische Akteure erhoffen, dürften diesen Effekt des Nebenbedingungsparadigmas hingegen als unwillkommene Begrenzung einschätzen.

Diese Einengung auf eine minimale Form von Akteurshaftigkeit – ob nun gewollt oder ungewollt – scheint aber nicht nur gegen verbreitete Intuitionen zu verstoßen, denen zufolge künstlich intelligente Systeme doch ein anderes, ein komplexeres Verhältnis zu unseren moralischen Normen haben. Es scheint auch nicht recht zur aufwändigen und differenzierten Debatte in der Maschinenethik zu passen. Es sollte sich also lohnen, nachzufragen, was für ein Schritt über das Nebenbedingungsparadigma hinaus erforderlich wäre, um davon sprechen zu können, dass KIs ethische Grenzen gesetzt werden.

8.6.1 Wie könnten der KI ethische Grenzen gesetzt werden?

Diese Frage lässt sich in mindestens zweierlei Hinsicht verstehen: Einerseits kann es die Frage danach sein, welche Eigenschaften und Fähigkeiten ein künstlich intelligentes System mindestens haben muss, damit die Rede davon sein kann, dass es Normen zustimmen kann. Typische Kandidaten dafür sind in der Debatte schnell identifiziert: exemplarisch phänomenales Bewusstsein³⁴, Intentionalität im Sinne von Gerichtetheit³⁵, Repräsentation³⁶, normative Commitments³⁷ oder Willensfreiheit³⁸.

So wichtig und spannend diese Debatte ist, soll sie hier nicht geführt werden. Denn die Frage nach der Grenzsetzung beinhaltet noch eine zweite Dimension, nämlich: Welchen Effekt muss ein künstlich intelligentes System auf die fraglichen Regeln haben können?

Diese Untersuchungsrichtung muss vielleicht noch ein wenig erläutert werden. Warum sollten wir fragen, ob irgendein Akteur oder eine Akteurin Einfluss auf eine Regel hat? Der Hintergrund besteht schlicht darin, dass die Rede davon, dass moralische Regeln auf die Zustimmung von Personen ausgerichtet sein müssen, leer bleibt, wenn Akteure und Akteurinnen nur die Möglichkeit haben, einer Regel zuzustimmen oder aufzuhören, an einer moralischen Gemeinschaft teilzuhaben. Es scheint erfor-

34 Vgl. Purves et al. (2015), Coeckelbergh (2010).

35 Vgl. Haugeland (1990).

36 Vgl. Searle (1980).

37 Vgl. Brandom (1994) und die Anwendung des Inferentialismus auf KI bei Heinrichs/Knell (2021).

38 Vgl. Hellström (2013).

derlich zu sein, dass moralische Akteure und Akteurinnen auch auf andere Weise auf moralische Regeln reagieren können. Und in der Tat können das prototypische moralische Akteure und Akteurinnen.

Moralische Akteure und Akteurinnen, wie wir sie bislang kennen, können sehr unterschiedliche Effekte auf moralische Regeln zeitigen, von denen hier nur ein paar Beispiele vorzuführen möglich ist. Sie können gravierenden Einfluss haben wie etwa, eine moralische Regel für eine ganze Gemeinschaft in Frage zu stellen. Das dürfte extrem selten der Fall sein, aber Beispiele dafür ließen sich bei Figuren wie Martin Luther King oder Rosa Parks finden. Sie können eine moralische Gemeinschaft zur Präzisierung herausfordern, wie das derzeit durch Vertreter:innen für Rechte von Trans-Personen geschieht. Sie können aber auch weniger augenfällige, aber dennoch relevante Effekte haben, indem sie beispielsweise durch das Einhalten oder Ablehnen einer Regel als Beispiel für Ihre Umgebung agieren oder durch öffentliches Einfordern von Regelkonformität eine Form sozialer Sanktion ausüben und damit zur sozialen Verbindlichkeit einer Regel beitragen oder ihr eben abträglich sind.

Die vielschichtige Frage nach dem Verhältnis von Regeln und Akteuren oder Akteurinnen, insbesondere nach deren Effekt auf Regeln steht generell seltener im Mittelpunkt der Aufmerksamkeit als diejenige nach der Verfasstheit moralischer Akteure und Akteurinnen. Dies gilt insbesondere für die Diskussion um künstliche Akteure. Um die Frage zu beantworten, welchen Effekt KIs auf Regeln zeitigen können müssen, um als moralische Akteure in Frage zu kommen, bedarf es einer kurzen Klärung, was hier mit ›Regel‹ gemeint ist.

Der Regelbegriff bezieht sich hier auf mehr als lediglich verhaltensinhärente Normen. Zur Erklärung: Eine Vielzahl von menschlichen Aktivitäten lässt sich mit der Spiel-Metapher beschreiben. Das ist auch für die Moral – wenn vielleicht auch als Grenzfall – möglich. Für viele solche Aktivitäten gilt, dass sie gespielt bzw. ausgeübt werden können, ohne dass die Teilnehmenden in der Lage sind, die Regeln zu formulieren³⁹. Die Regeln sind ihrem Handeln inhärent. Für einige Spiele gilt wahrscheinlich, dass ihre Regeln *nur* verhaltensinhärent vorliegen. Für den gegenwärtigen Kontext, d. h. das Spiel ›Moral‹, ist das aber nicht hinreichend. Wenn der Anspruch moralischer Normen wirklich ist, dass Personen ihnen zustimmen können müssen, dann reicht eine nur verhaltensinhärente Norm nicht aus. Es bedarf irgendeines intentionalen Objekts der Zustimmung, und das besteht nicht in einer Menge von Verhaltensweisen, sondern in einer – noch so vagen – Formulierung der sie anleitenden Regel.

Dagegen ließe sich einwenden, dass eine Person sehr wohl in der Lage sei, den Regeln einer Aktivität ihre Zustimmung zu geben, wenn sie deren Regeln nicht formulieren kann. Sie müsse lediglich weiterhin an der Aktivität teilnehmen. Das ist aber aus dreierlei Gründen in diesem Kontext nicht hinreichend. Erstens ist es, wie oben beschrieben, kaum möglich, das Spiel der Moral ganz zu verlassen, ohne sich der Bedingungen eines gelingenden oder überhaupt eines Lebens zu berauben. Zustimmungsverweigerung ist also keine echte Option. Zweitens gerät auf diese Weise

39 Vgl. Sellars (1954).

die Möglichkeit aus dem Blick, sich zu einzelnen Regeln zu verhalten. Weiterspielen oder Aufhören ist ein Verhältnis zum Spiel als Ganzem, nicht zu einzelnen Regeln. Und drittens, direkt damit verbunden, gerät die Möglichkeit der Revision innerhalb des Spiels dadurch aus dem Blick. Gäbe man sich mit Weiterspielen als Zustimmung zufrieden, dann verlöre man die Sensitivität dafür, dass Personen – oder Handelnde – eben nicht nur in der Lage sind, allen Regeln einfach zu folgen, sondern dass sie zumindest in vielen Fällen in der Lage sind, einzelne Regeln zu verweigern, sie zu revidieren oder sie für besonders wichtig zu erachten, ohne das Spiel insgesamt zu verlassen.

Kurzum, für den gegenwärtigen Kontext wird unter ›Regel‹ eine Formulierung von Normen verstanden, die auch im Verhalten von Akteuren und Akteurinnen präsent sind. Natürlich können nicht alle moralischen Akteure und Akteurinnen alle Regeln des Spiels ›Moral‹ korrekt formulieren. Wahrscheinlich gibt es noch nicht einmal ein vollständiges Set von Regeln dieses Spiels. Um zu thematisieren, welchen Effekt KI-Systeme auf moralische Regeln haben, muss aber mindestens eine grobe Formulierbarkeit einiger dieser Regeln vorausgesetzt werden.

Weil mit dieser Bestimmung des Regelbegriffs sowohl auf die Formulierung als auch auf die Präsenz in Verhalten referiert wird, reicht es für einen Effekt auf eine Regel nicht aus, dass sie grammatisch umgeformt wird. Ein Effekt liegt nur dann vor, wenn nicht nur die Formulierung der Norm, sondern auch deren Präsenz im Verhalten von Akteuren und Akteurinnen betroffen ist. Sollte beispielsweise ChatGPT eine clevere bedeutungserhaltende Umformulierung der dafür geltenden Regeln generieren können, dann zählt das nicht als ein relevanter Effekt von KIs auf moralische Regeln.

Davon ist bereits ein möglicher interessanter Effekt abzugrenzen: Sollte eine bedeutungserhaltende Reformulierung dazu führen, dass eine bestehende Regel von anderen Akteuren und Akteurinnen besser verstanden und deshalb vermehrt angewandt, übernommen, verworfen oder revidiert wird, dann läge sehr wohl bereits ein Effekt eines KI-Systems auf moralische Regeln vor.

8.6.2 Welchen Effekt können künstlich intelligente Systeme auf moralische Regeln haben?

Es kann hier es nicht darum gehen, alle möglichen, nicht einmal alle aktuellen Effekte künstlich intelligenter Systeme auf moralische Normen aufzuführen. Vielmehr werden im Folgenden zwei Sachverhalte aufgezeigt, nämlich erstens, dass das Nebenbedingungsparadigma – so sinnvoll es aus anderen Gründen auch sein mag – das zentrale Hemmnis für weitere Effekte von KIs auf moralische Regeln ist. Zweitens soll vorgeführt werden, dass bereits jetzt komplexere Effekte möglich sind, als im Nebenbedingungsparadigma eigentlich angelegt sind.

Zunächst: Wie oben beschrieben ist das Nebenbedingungsparadigma dergestalt, dass es das Verhalten künstlich intelligenter Systeme an den Vorgaben moralischer Regeln ausrichtet. Obwohl es im Prinzip darin so etwas wie das Abwägen von Regeln

geben könnte, wären doch die Regeln für die Abwägung wiederum vorgegeben, eine Evaluation der Regeln durch die KI ist nicht vorgesehen.

Dennoch lassen sich bereits an gegenwärtigen KIs einige Effekte auf Normen beobachten. Damit ist nicht einfach gemeint, dass wir unsere Werthaltungen oder Normen anpassen, wenn Erfolge der KI-Forschung etablierte Überzeugungen zur Einzigartigkeit oder Komplexität einer Tätigkeit revidieren. Solche Effekte sind relativ verbreitet, wie etwa ein Paradoxon der KI-Forschung: Sobald eine Tätigkeit – wie Schach oder radiologische Diagnosestellung – durch künstliche Intelligenz realisiert wird, gilt sie einerseits plötzlich nicht mehr in dem vorherigen Maße als intelligent, auch wenn sie zuvor als klassisches Beispiel dafür gehandelt wurde. Andererseits scheint auch die Wertschätzung für die jeweiligen Tätigkeiten unter der Automatisierung zu leiden. Dabei handelt es sich aber lediglich um eine weitgehend ungerechtfertigte Veränderung oberflächlicher Evaluationen durch die Möglichkeit von Automatisierung, nicht um einen systematischen Effekt konkreter KI-Systeme oder Techniken auf etablierte moralische Normen.

Sehr viel interessanter ist ein von Anderson und Anderson in ihrem Programm MedEthEx realisierter Effekt. Sie haben ein KI-System mit einem Datensatz weithin akzeptierter Lösungen ethischer Dilemmata zwischen Prima-facie-Pflichten trainiert und auf dieser Basis versucht, Entscheidungsprinzipien zu identifizieren. Dabei sei in der Tat ein Entscheidungsprinzip explizit gemacht worden, das den Lösungen im Trainingsdatensatz zwar inhärent, aber eben nicht explizit war: »A health-care worker should challenge a patient's decision if it isn't fully autonomous and there's either any violation of nonmaleficence or a severe violation of beneficence.«⁴⁰ Das Programm greift also auf Verhaltensdaten und Formulierungen von Regeln, nämlich Prima-facie-Pflichten, zurück und macht etwas explizit, was den Entscheidungsträger:innen in dieser Form in vielen Fällen nicht bewusst gewesen sein dürfte.

Damit generiert dieses Programm – wie auch einige andere ähnlich gelagerte Systeme⁴¹ – die Möglichkeit für die moralische Gemeinschaft, sich zu dieser impliziten Norm neu zu verhalten, sie zu prüfen und zu revidieren, auch wenn das Programm selbst nichts dergleichen tut. Es handelt sich also um einen Grenzfall von Einfluss der KI. Zwar wird eine moralische Regel und das Verhalten von Akteuren und Akteurinnen zu dieser moralischen Regel explizit repräsentiert und das Verhältnis unterschiedlicher Regeln zueinander analysiert. Es handelt sich aber um einen Grenzfall, weil es bei einer Deskription bleibt. Das System betreibt so etwas wie deskriptive Ethik. Allein das dürfte in hermeneutischer Hinsicht schon interessant sein, denn welche Beschreibung die KI ausgibt, scheint durch den Beschreibungsgegenstand allein nicht festgelegt, sondern von der Trainingsgeschichte des Systems mitbestimmt zu sein. Diese Form der Explikation moralischer Regeln macht die KI also bereits zu einer Art epistemischen Als-ob-Akteur.⁴² Unter einem Als-ob-Akteur verstehe ich in diesem Kontext ein System, das die für Akteure und Akteurinnen typischen Leistun-

40 Anderson/Anderson (2007): 23.

41 Vgl. Cervantes et al. (2020).

42 Ich danke Ulrich Steckmann für diesen Hinweis.

gen in einem Handlungsfeld erbringt, ohne dass aber klar wäre, dass es die typischen Eigenschaften von Akteuren und Akteurinnen in diesem Feld – etwa Intentionalität oder Absichtlichkeit – aufweist⁴³. Grundsätzlich scheint damit aber das Potential aufgezeigt, eine andere als eine schlicht deskriptive Beziehung zu den repräsentierten Normen zu generieren.

Dieses Potential ist deshalb gegeben, weil neben der Fähigkeit zur Repräsentation von Regeln und Verhalten weitere Komponenten der Fähigkeit, sich zu Normen zu verhalten, bereits klarerweise zum Repertoire von KI-Systemen gehören. So sind etwa die Fähigkeiten zu Konsistenzprüfung, Kohärenzprüfung⁴⁴, kontrafaktische Konstruktionen⁴⁵ und *value-based reasoning*⁴⁶, also der Prüfung des Beitrages einer Norm zur Realisierung eines Ziels schon realisiert worden. Diese Optionen lassen sich grundsätzlich auf einmal repräsentierte Systeme moralischer Regeln anwenden, um einen Beitrag zu deren Evaluation zu leisten. Obwohl es zweifelsfrei noch zahlreiche Hürden dafür gibt, dass künstlich intelligente Systeme moralische Regelsysteme zum Thema ihrer Informationsverarbeitung machen, statt sie lediglich als Nebenbedingungen zu verwenden, liegt eine Grundausrüstung, derer es dafür bedürfte, also bereits vor.

Gegenwärtig dürfte aber – gerade aufgrund des oben beschriebenen Nebenbedingungsparadigmas – die Entwicklung von KI-Systemen, die moralische Regeln auf ihre Konsistenz und ihre Kohärenz miteinander prüfen, eher eine Ausnahme bilden. Das gilt umso mehr für Systeme, die überprüfen, ob konkrete, moralische Regeln geeignet oder gar optimal dafür sind, bestimmte Werte zu realisieren. Selbst sogenannte künstliche moralische Berater – im noch fiktionalen Fall von Giubilini und Savulescu⁴⁷ ebenso wie im realen von Anderson und Anderson⁴⁸ oder Meier⁴⁹ – beschränken sich typischerweise darauf, Rat für Einzelhandlungen innerhalb eines vorgefertigten und durch das System nicht weiter befragten Gerüsts moralischer Regeln zu geben. Auch diese Systeme sind bislang nicht darauf ausgelegt, moralische Regeln zu revidieren oder weiterzuentwickeln, oder wenigstens Vollmitgliedern der moralischen Gemeinschaft einen epistemischen Anlass für solch eine Revision zu geben.

8.6.3 Welchen Effekt dürfen KIs auf moralische Regeln haben?

Ich hoffe plausibel gemacht zu haben, dass ein verbreitetes Paradigma der Maschinenethik, das Nebenbedingungsparadigma, darauf ausgelegt ist, das Verhalten von KIs den Vorgaben moralischer Regeln anzupassen, nicht aber moralische Akteure in

43 Vgl. Johansson (2010).

44 Vgl. Suwa et al. (1982).

45 Vgl. Pereira/Saptawijaya (2016): 81 ff.

46 Vgl. Badea (2022).

47 Vgl. Giubilini/Savulescu (2018).

48 Vgl. Anderson/Anderson (2007).

49 Vgl. Meier et al. (2022).

einem starken Sinn zu generieren. Diese Ausrichtung läuft zwar dem verkündeten Selbstverständnis von Maschinenethiker:innen oft zuwider. Das Nebenbedingungsparadigma ist immerhin fester Bestandteil des Versuchs sogenannte künstliche moralische Akteure zu schaffen. Es entspricht damit aber zugleich dem impliziten Selbstverständnis von Maschinenethiker:innen. Diese künstlich moralischen Akteure werden nur deshalb geschaffen, weil man trotz steigender Komplexität des Verhaltens *garantieren* möchte, dass KI-Systeme sich in einer Art verhalten, die moralischen Vorgaben genügt.

Es hat sich aber auch gezeigt, dass schon jetzt Systeme existieren, die moralische Normen nicht nur schlicht einhalten, sondern diese explizieren, repräsentieren und zueinander in Beziehung setzen. Es sind zudem bereits jetzt Systeme möglich, die über diese deskriptiven Tätigkeiten hinausgehen und moralische Normen partiell überprüfen. Damit ist die Möglichkeit eröffnet, dass sich KIs in anderer Weise zu moralischen Grenzen verhalten als vom Nebenbedingungsparadigma vorgesehen. Sie können mindestens die Möglichkeit zur Überprüfung und Kritik moralischer Regeln generieren.

Wie aber oben gezeigt ist das Nebenbedingungsparadigma durchaus gerechtfertigt. Es soll sicherstellen, dass dem Verhalten von KIs Grenzen gesetzt sind, wenn auch eben nicht in der Form, wie man moralischen Akteuren und Akteurinnen Grenzen setzt. Man wird also fragen müssen, ob die Alternative, nämlich KIs einen Effekt auf moralische Regeln zu erlauben, ebenfalls gerechtfertigt ist. Diese Frage lässt sich weder pauschal noch kurz beantworten.

Dennoch möchte ich hier noch ein Bedenken zerstreuen, das uns darin hindert, die Fähigkeiten zur Revision moralischer Regeln zu nutzen, die künstlich intelligente Systeme möglicherweise haben. Das Bedenken, das sich direkt aus dem Nebenbedingungsparadigma der Maschinenethik ergibt, ist folgendes: Müssen wir nicht fürchten, dass KIs moralische Regeln, die wir ihnen auferlegen wollen, aufgrund von uns opaken Prozessen verwerfen und aufhören, sich daran zu halten? Würde damit nicht das Ziel der ganzen Übung hintertrieben, künstliche moralische Akteure zu erschaffen?

Dieses Bedenken ist sicherlich gerechtfertigt, insofern es tatsächlich sein könnte, dass die kritische Prüfung unserer moralischen Regeln und unseres Verhaltens dazu erheblichen Revisionsbedarf erzeugt. Daraus aber ein Szenario zu folgern, in der KI sich diesen Regeln verweigert, dürfte voreilig sein. Der Umstand allein, dass ein System eine Regel nach rationaler Prüfung verwerfen würde, versetzt es nicht schon in die Lage, die Regel effektiv abzulehnen. Das gilt bereits für Menschen und erklärt, warum unsere moralischen Rahmenbedingungen zuweilen stabiler sind als der überlappende gesellschaftliche Konsens, der sie stützt. Es gilt umso mehr für künstlich intelligente Systeme, die nicht einmal sanktioniert werden müssen, sondern programmiert werden können, moralische Regeln einzuhalten. Wie oben erwähnt wird regelmäßig diskutiert, welche Eigenschaften KIs haben müssten, um davon sprechen zu können, dass sie einer Regel zustimmen. Und in dieser Debatte wird fast genauso häufig konstatiert, dass derzeit und bis auf absehbare Zeit kein KI-System über diese Eigenschaften verfügt. Sieht man einmal von hochspekulativen Szenarien über Su-

perintelligenzen ab, besteht also kaum das Risiko, dass KIs jenseits des Nebenbedingungsparadigmas sich plötzlich gegen unsere Regeln verhalten. Es besteht höchstens das Risiko, dass sie uns die Unzulänglichkeit dieser Regeln vorführen.⁵⁰

8.7 Literatur

- Anderson, Michael/Anderson, Susan Leigh (2007): *Machine ethics: Creating an ethical intelligent agent*, in: *AI Magazine* 28/4, 15–26.
- Anderson, Michael/Anderson, Susan Leigh (2011): *Machine Ethics*. Cambridge University Press: Cambridge.
- Badea, Cosmin (2022): *Have a break from making decisions, have a MARS: The Multi-valued Action Reasoning System*, in: Bramer, Max/Stahl, Frederic (Hg.): *Artificial Intelligence XXXIX: 42nd SGAI International Conference on Artificial Intelligence, AI 2022*, Springer: Cham, 359–366.
- Behdadi, Dorna/Munthe, Christian (2020): *A Normative Approach to Artificial Moral Agency*, in: *Minds and Machines* 30/2, 195–218.
- Brandom, Robert (1994): *Making it explicit. Reasoning, representing, and discursive commitment*, Harvard University Press, Cambridge (Mass.)
- Cervantes, José-Antonio/López, Sonia/Rodríguez, Luis-Felipe/Cervantes, Salvador/Cervantes, Francisco/Ramos, Félix (2020): *Artificial Moral Agents: A Survey of the Current Status*, in: *Science and Engineering Ethics* 26/2, 501–532.
- Coeckelbergh, Mark (2010): *Moral appearances: emotions, robots, and human morality*, in: *Ethics and Information Technology* 12/3, 235–241.
- Danaher, John (2016): *Robots, law and the retribution gap*, in: *Ethics and Information Technology* 18/4, 299–309.
- Floridi, Luciano/Sanders, J. W. (2004): *On the Morality of Artificial Agents*, in: *Minds and Machines* 14/3, 349–379.
- Ganguli, Deep et al. (2023): *The Capacity for Moral Self-Correction in Large Language Models*, arXiv.2302.07459.
- Giubilini, Alberto/Savulescu, Julian (2018): *The Artificial Moral Advisor. The »Ideal Observer« Meets Artificial Intelligence*, in: *Philosophy & Technology* 31/2, 169–188.
- Haugeland, John (1990): *The Intentionality All-Stars*, in: *Philosophical Perspectives* 4, 383–427.
- Heinrichs, Bert/Heinrichs, Jan-Hendrik/Rüther, Markus (2022): *Künstliche Intelligenz*, de Gruyter, Berlin/New York.
- Heinrichs, Bert/Knell, Sebastian (2021): *Aliens in the Space of Reasons? On the Interaction Between Humans and Artificial Intelligent Agents*, in: *Philosophy & Technology* 34/4, 1569–1580.
- Hellström, Thomas (2013): *On the moral responsibility of military robots*, in: *Ethics and Information Technology* 15/2, 99–107.
- Hume, David (1987): *Essays, moral, political, and literary*, LibertyClassics, Indianapolis.
- Jobin, Anna/Ienca, Marcello/Vayena, Effy (2019): *The global landscape of AI ethics guidelines*, in: *Nature Machine Intelligence* 1/9, 389–399.
- Johansson, Linda (2010): *The Functional Morality of Robots*, in: *International Journal of Technoethics* 1/4, 65–73.
- Johnson, Deborah G. (2006): *Computer systems: Moral entities but not moral agents*, in: *Ethics and Information Technology* 8/4, 195–204.

50 Und um einmal ganz spekulativ zu sein: Sollten die Szenarien zu Superintelligenzen und deren zukünftiger Überlegenheit irgendwann zutreffen, dann wäre es eventuell besser, bis dahin in Aushandlung gestanden zu haben, statt in einer Form der Unterwerfung.

- Liu, Yuxin/Moore, Adam/Webb, Jamie/Vallor, Shannon (2022): *Artificial Moral Advisors: A New Perspective from Moral Psychology*, in: AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, 436–445.
- Lundgren, Björn (2023): *In defense of ethical guidelines*, in: AI and Ethics 3/3, 1013–1020.
- Matthias, Andreas (2004): *The responsibility gap: Ascribing responsibility for the actions of learning automata*, in: Ethics and Information Technology 6/3, 175–183.
- Meier, Lukas J./Hein, Alice/Diepold, Klaus/Buyx, Alena (2022): *Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept*, in: The American Journal of Bioethics 22/7, 4–20.
- Miller, Keith W./Taddeo, Mariarosaria (Hg.) (2017): *The Ethics of Information Technologies*, Routledge, London.
- Misselhorn, Catrin (2018): *Grundfragen der Maschinenethik*, Reclam, Stuttgart.
- Moor, James H. (2006): *The Nature, Importance, and Difficulty of Machine Ethics*, in: IEEE Intelligent Systems 21/4, 18–21.
- Pereira, Luís Moniz/Saptawijaya, Ari (2016): *Programming Machine Ethics*, Springer, Cham.
- Purves, Duncan/Jenkins, Ryan/Strawser, Bradley J. (2015): *Autonomous Machines, Moral Judgment, and Acting for the Right Reasons*, in: Ethical Theory and Moral Practice 18/4, 851–872.
- Scanlon, Thomas (1998): *What we owe to each other*, Belknap Press, Cambridge (Mass.)
- Searle, John R. (1980): *Minds, Brains, and Programs*, in: Behavioral and Brain Sciences 3/3, 417–425.
- Sellars, Wilfrid (1954): *Some reflections on language games*, in: Philosophy of Science 21/3, 204–228.
- Sullins, John P. (2006): *When is a robot a moral agent*, in: International Review of Information Ethics 6/12, 23–30.
- Suwa, Motoi/Scott, A. Carlisle/Shortliffe, Edward H. (1982): *An approach to verifying completeness and consistency in a rule-based expert system*, in: AI Magazine 3/4, 16.
- Tigard, Daniel W., (2021): *There Is No Techno-Responsibility Gap*, in: Philosophy & Technology 34/3, 589–607.
- Turing, Alan M. (1950): *Computing Machinery and Intelligence*, in: Mind 59/236, 433–460.
- Turing, Alan M. (1951): *Intelligent machinery, a heretical theory. A lecture given to '51 Society' at Manchester*, The Turing Archive, Manchester.
- Van Wynsberghe, Aimee/Robbins, Scott (2019): *Critiquing the Reasons for Making Artificial Moral Agents*, in: Science and Engineering Ethics 25/3, 719–735.
- Wallach, Wendell/Allen, Colin (2009): *Moral machines. Teaching robots right from wrong*, Oxford University Press, Oxford/New York.
- Whitby, Blay (2008): *Computing machinery and morality*, in: AI & Society 22/4, 551–563.